DOCUMENT RESUME

ED 135 826 TM 006 051

AUTHOR Abranovic, Wynn A.
TITLE Test Types and the Relation of Test Scores to Knowledge.
NOTE 13p.

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS *Guessing (Tests); *Knowledge Level; Mathematical Models; Multiple Choice Tests; *Objective Tests; Probability Theory; Response Style (Tests); Scores; *Statistical Analysis
IDENTIFIERS Matching Tests; True False Tests

ABSTRACT

Two basic issues are discussed that involve the Carver model of the relationship between test score and knowledge. The first of these concerns different types of tests which have different probabilities of obtaining the correct answer due to guessing. A derivation is shown which places different test types "on-a-par"--meaning that two tests are designed to maintain a comparable spread in scores; more specifically, this is defined such that two tests have approximately the same rate of change in test score as the proportion of knowledge changes. An equation is provided to achieve this which compares a true-false test to multiple choice or matching tests. It is concluded that for two tests to be "on-a-par" the number of questions must differ, given the probabilities of obtaining correct answers by guessing differ. The second issue discussed involves approximate confidence intervals on the proportion of knowledge obtained. An equation is provided which gives an approximation for the standard deviation of the proportion of knowledge; limits for three standard deviations appear in another equation. This connection of test scores and knowledge level is very important. It provides a standard basis for evaluating performance. In addition, it provides some insight into teaching effectiveness. (BC)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED135826

TM

TEST TYPES AND THE RELATION OF TEST SCORES TO KNOWLEDGE

Wynn A. Abranovic--University of Massachusetts

INTRODUCTION

Throughout the world, colleges and universities all use a variety of tests and examinations. There are entrance, aptitude, admissions, achievement, advancement and psychological tests, that form an integral part of educational systems. Excluding essay examinations there are three prevalent examination structures that are commonly used: true-false, multiple-choice and matching. Each form has advantages and disadvantages which make it more or less popular than the others. Many prefer true-false tests rather than multiple-choice because true-false are generally easier to construct and grade. From my experience matching tests take a back seat in preference to the other two, although matching tests are easy to make up, simple to grade, and very resistant to the effects of guessing.

It is well known that different tests are not equally efficient. For example, a true-false test usually provides the opportunity for more correct answers due to guessing than does a multiple-choice test. The proof of this is based on probability theory. The theory leads naturally to an inquiry concerning the number of questions needed on different types of tests. If one test is more efficient than another, should the more efficient test require fewer questions to achieve the same objective measure of knowledge? This paper argues

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM006 051

that the answer is yes. In fact, both the probability theory and Carver's (1969), surprisingly, point to the same conclusion and an approximate relationship is found. This paper also discusses approximate confidence bounds on the proportion of knowledge obtained.

A REVIEW OF THE CARVER MODEL

Carver (1969, p. 64) develops the following model relating the number of questions answered correctly, x, to the knowledge, y, known to the student:

$$x = p_i n_i + \frac{2 q_i n_i y}{N} - \frac{q_i n_i y^2}{N^2} \qquad (1)$$

where,

x = the number of questions answered correctly on an examination.

n_i = the number of questions on examination i.

p_i = the probability of answering a question correctly by guessing on test i (i.e., $p = \frac{1}{2}$ for true-false; $p = \frac{1}{4}$ for a multiple choice examination with four choices choices per question).

N = the total number of units of knowledge covered by the course.

y = the number of units of knowledge covered by the course that is known to the student.

$q_i = 1 - p_i$.

If the model is rewritten, it becomes an easier one to use. Let $y/N = z$, where z is the proportion of N that is y; in words, the proportion of total knowledge obtained, say, in a

course of study. Clearly, z is expected to be greater than or equal to zero, no knowledge, and less than or equal to one, perfect knowledge. Now, equation (1) can be rewritten as

$$x = p_i n_i + 2q_i n_i z - q_i n_i z^2 \qquad (2)$$

Let's consider a simple application of (2) in order to obtain a better understanding of its behavior. Suppose we consider a one-hundred question, true-false test: this implies that $n = 100$ and $p = q = \frac{1}{2}$, that is, the probability of getting an answer correct by guessing is $p = \frac{1}{2}$. When $z = 0$, no knowledge,

$$x = pn = \tfrac{1}{2}(100) = 50.$$

One can expect on the average to answer fifty questions correctly by guessing. This is point A on Exhibit 1. This exhibit is a plot of the number of questions answered correctly on an examination versus the proportion of total knowledge obtained. When $z = \frac{1}{2}$, one-half of the knowledge is obtained and

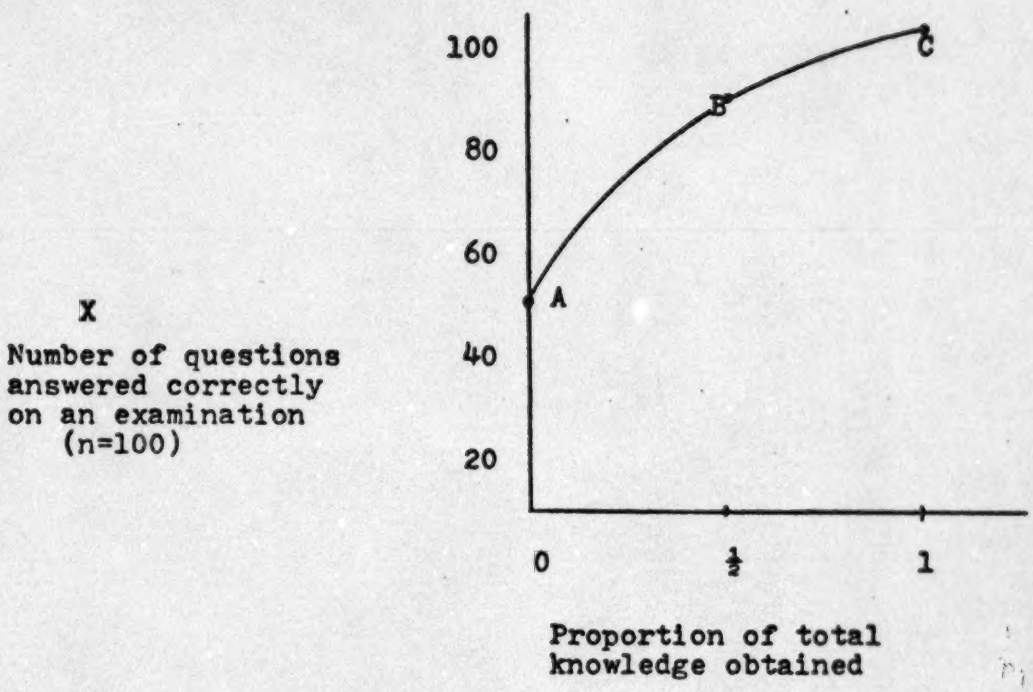
$$x = 50 + (2)(\tfrac{1}{2})(100)(\tfrac{1}{2}) - (\tfrac{1}{2})(100)(\tfrac{1}{2})^2 = 87.5$$

"Place Exhibit 1 about here."

Thus, by guessing and knowing one-half of knowledge, one can expect, based on this model, to answer 87.5 questions correctly. This is point B in Exhibit 1. When $z = 1$, all the knowledge is known and one answers all the questions correctly. This is point C on the graph. A smooth curve, ABC, is drawn, at other levels of knowledge.

Any test other than a true-false type, with the same number of questions changes the starting point A but not the ending point C. For both multiple-choice (with more than two

Exhibit 1


X
Number of questions answered correctly on an examination (n=100)
100
80
60
40
20
A
B
C
0
½
1
Proportion of total knowledge obtained

choices per question) and matching tests, the starting point A is lowered, with the lower starting point for matching. This means that guessing is less effective in obtaining correct answers when multiple-choice or matching is used in place of true-false.

It is also possible to change the test type and the number of questions on the test. The curve A,B,C may be shifted up or down with simultaneous changes in the tilt and curvature. This changes starting, ending and intermediate points. If n is less than 100, the curve drops and, if n is greater than 100, the curve shifts upward (at the same time the tilt and curvature may change, but this depends on p and q).

Carver uses the quadratic form (1) to search for the best indicant of the amount learned in classroom situations. It is concluded that the final examination is the best indicant of amount learned. Carver's model will be used here for a different purpose; it will form a basis for comparing test types (true-false, multiple-choice, and matching).

Because it is well known that test types differ in their ability to deal with the inflationary pressures of guessing, this paper suggests a way of making different tests somewhat comparable. If one is examining several groups of individuals, he is often in a situation in which different tests must be made for each group. Sometimes a true-false test is given in one group and a multiple-choice test is given in another. Also, some tests are constructed so that they contain a mixture of true-false, multiple choice and matching questions; it is

desirable that parts of the mixture are in balance. This paper shows the number of questions needed to achieve balance among the various test types. The criterion used here can be stated as follows:

Make $x_{ib} - x_{iw}$ constant on the average for different test types, i, where

x_{ib} = the number of questions answered by the best performer at knowledge level z_b on test i.

x_{iw} = the number of questions answered by the worst performer at knowledge level z_w on test i.

As an example, suppose one has either a 100 question true-false test or a 50 question multiple-choice test and $x_{ib} - x_{iw} = 20$. If the best student is expected to get a perfect paper, one would expect the worst student on the true-false test to get an 80; on the multiple-choice test, one would expect the worst student to get a 30. On either test, one can easily partition the twenty point range for establishing letter grades. To establish a constant spread, $x_{ib} - x_{iw}$, partial derivatives are used which give the rates of change in test scores as the knowledge level changes.

The partial derivative of x with respect to z in equation (2), gives the Rate Of Change in test scores as a function of Knowledge (ROCK). ROCK shows how many units the test score changes with each unit change in knowledge. Under the criterion adopted here, it is desirable to have tests with equal ROCK'S; this implies that the change in test scores relative to the knowledge will be approximately the same even though the absolute

score magnitudes may differ.

If there are n_1 true-false questions in test one and n_2 (four choices per question) multiple-choice questions in test two, then the ROCK'S, theoretically, cannot be equal when each test has the same number of questions. This means that one of the tests must have more questions than the other in order to maintain a constant change in scores relative to the knowledge. An approximate relationship between n_1 and n_2 needed to obtain a constant rate of change will be derived via the Carver model in the following section. Confidence limits on the proportion of knowledge obtained will also be discussed.

METHODOLOGY AND APPLICATIONS

The partial derivative of equation (2) with respect to knowledge at z_{ib} for test type #1 yields

$$\text{ROCK}\ (1)_b = \frac{dx^{(1)}}{dz_{ib}} = 2q_1n_1 - 2q_1n_1z_{1b} \qquad (3)$$

at knowledge level z_{iw}

$$\text{ROCK}\ (1)_w = \frac{dx^{(1)}}{dz_{iw}} = 2q_1n_1 - 2q_1n_1z_{1w} \qquad (4)$$

The average of (3) and (4) is

$$q_1n_1[2 - (z_{1b} + z_{1w})] \qquad (5)$$

For test type # 2, a similar process yields

$$q_2n_2[2 - (z_{2b} + z_{2w})] \qquad (6)$$

Our criterion requires (5) and (6) to be equal, and results in

$$q_1n_1 = q_2n_2 \qquad (7)$$

$q_1 = \frac{1}{2}$ for a true-false test, so the number of questions needed on the other test types is given by

$$n_2 = \frac{n_1}{2q_2} \qquad (8)$$

Equation (8) indicates that any test (where $1 \geq q_2 > \frac{1}{2}$, that is not true-false) gives, on the average, the same relative change in scores per unit change in knowledge with fewer questions, i.e., $n_2 < n_1$. As an application, suppose a true-false test is compared to a multiple choice test with four choices per question; $q_2 = 3/4$. From (8) $n_2 = 2/3\ n_1$.

This implies that the discriminatory power of each test is approximately equal but the multiple choice test needs only two-thirds as many questions. For a matching test, q is even closer to 1 and the test requires fewer questions.

Approximately confidence limits can be constructed on the proportion of total knowledge obtained by the group being examined.

Let:

$\bar{x}$ = the average score of the group being tested.

σ_x = the standard deviation of the scores.

$\bar{z}$ = the average proportion of total knowledge obtained by the group.

σ_z = the standard deviation of the proportion of total knowledge obtained by the group.

$\bar{z}$ can be estimated from Exhibit 1 by finding the value of z that corresponds to $\bar{x}$. As an example, suppose $\bar{x} = 87.5$ then an estimate of $\bar{z} = .50$. Estimating σ_z is a little more difficult. We make use of a Taylor series approximation technique, which is well known; it is demonstrated in [6, p. 152]. Let us write equation (2) as

$$g(z) = x = p_i n_i + 2q_i n_i z - q_i n_i z^2 . \qquad (9)$$

where $g(\cdot)$ is a function of z. From [6, p. 152] we can write

$$\sigma^2_{g(z)} \doteq g'[(z = \bar{z})]^2 \sigma_z^2. \tag{10}$$

The derivative of (9) is

$$g'(z = \bar{z}) = 2q_i n_i - 2q_i n_i \bar{z}. \tag{11}$$

Then,

$$\sigma^2_{g(z)} \doteq [2q_i n_i - 2q_i n_i \bar{z}]^2 \sigma^2_z. \tag{12}$$

$$\sigma^2_{g(z)} \doteq 4(\sigma_z q_i n_i)^2 (1 - \bar{z})^2 \tag{13}$$

Since,

$$\sigma_x = \sqrt{\sigma^2_{g(z)}} \doteq 2\sigma_z q_i n_i (1-\bar{z}). \tag{14}$$

From (14), solving for σ_z,

$$\sigma_z \doteq \sigma_x / (2q_i n_i (1-\bar{z})). \tag{15}$$

When the normality assumption holds, confidence limits can be constructed on the proportion of total knowledge obtained. They are

$$\bar{z} \pm 3\sigma_z, \tag{16}$$

for limits of three standard deviations.

Consider the following application for a true-false examination. Let the average score be $\bar{x} = 87.5$, which is expected when fifty percent of the knowledge is obtained, as shown in Exhibit 1.

Suppose the standard deviation of the scores, σ_x, is two and one-half. This gives a range in scores of about eighty to ninty-five on a one-hundred point examination. From equation (15)

$$\sigma_z \doteq 2.5 / (2(\tfrac{1}{2}) 100 (\tfrac{1}{2})) = .05. \tag{17}$$

Employing this result in (16), one obtains

$$.50 \pm 3(.05) \tag{18}$$

Thus, the confidence limits on the proportion of total knowledge obtained by the group range from thirty-five to sixty-five percent of the knowledge. This is based on the results of a one-hundred point true-false examination.

DISCUSSION AND CONCLUSIONS

This paper discusses two basic issues that involve the Carver model of the relationship between test score and knowledge. The first of these, concerns different types of tests which have different probabilities of obtaining the correct answer due to guessing. A derivation is shown which places different test types "on-a-par". "On-a-par" in this case means that two tests are designed to maintain a comparable spread in scores; more specifically, this is defined such that two tests have approximately the same rate of change in test score as the proportion of knowledge changes. To achieve this, equation (8) can be used. This equation compares a true-false test to multiple-choice or matching tests. Using equation (7) any two test types can be compared. It can be concluded that for two tests to be "on-a-par" the number of questions must differ, given the probabilities of obtaining correct answers by guessing differ.

If test efficiency is indicated by the number of questions needed to obtain an estimate of the proportion of knowledge, matching-tests are most efficient and lie on one end of the efficiency continium while true-false tests fall on the other end and are least efficient. Multiple-choice tests lie somewhere in between. In the limit, q_2 approaches one and the

number of questions needed in test two approaches one-half of n_1 .

The second major issue discussed in this paper involves approximate confidence intervals on the proportion of knowledge obtained. Equation (15) gives an approximation for the standard deviation of the proportion of knowledge; limits for three standard deviations appear in equation (16). This connection of test scores and knowledge level is very important. It provides a standard basis for evaluating performance; in addition, it provides some insight into teaching effectiveness. Many other uses also are apparent.

REFERENCES

[1] Carver, Ronald P. "A Model for Using the Final Examination as a Measure of the Amount Learned in Classroom Learning". Journal of Educational Measurement, 2 (1969), pp. 59-68.

[2] Hambleton, R.K., et al. "A Comparison of the Reliability and Validity of Two Methods for Assessing Partial Knowledge on a Multiple-Choice Test". Journal of Educational Measurement, 2 (1970), pp. 75-90.

[3] Knapp, Thomas R. "A Note Concerning Carver's...Model..." Journal of Educational Measurement, 1 (1970), p. 51.

[4] Lord, F.M. and Novick, M.R. Statistical Theories of Mental Test Scores. Reading, Massachusetts: Addison-Wesley, 1968.

[5] Mosteller, et al. Probability with Statistical Applications. Reading, Massachusetts: Addison-Wesley, 1970.

[6] Papoulis, Athanasios. Probability, Random Variables, and Stochastic Processes, New York, New York: McGraw-Hill, 1965.

[7] Shuford, E.H. "How to Shorten a Test and Increase its Reliability and Validity". Technical Report SMC R-9, Shuford-Massengil Corporation. Lexington, Mass., May 1967.